Google AI

The 3rd YouTube-8M Large-Scale Video Understanding Workshop

October 28, 2018

ICCV 2019
Seoul, Korea

Agenda (Morning)

Time	Content	Presenter
9:00 - 9:05	Opening Remarks	Paul Natsev
9:05 - 9:20	**Overview of 2019 YouTube-8M Dataset & Challenge**	Joonseok Lee
Session 1		
9:20 - 9:50	**Invited Talk 1**: Human action recognition and the Kinetics dataset	Jitendra Malik
9:50 - 10:20	**Invited Talk 2**: Learning from Narrated Videos	Jean-Baptiste Alayrac
10:20 - 10:40	*Coffee Break*	
Session 2		
10:40 - 11:00	MediaPipe: A framework for building perception pipelines	Chris McClanahan
11:00 - 12:00	**Oral Session 1** ● Logistic Regression is Still Alive and Effective:The 3rd YouTube 8M challenge solution of the IVUL-KAUST team ● Multi-attention Networks for Temporal Localization of Video-level Labels ● A segment-level classification solution to the 3rd YouTube-8M Video Understanding Challenge	● IVUL-KAUST (#11) ● Locust (#13) ● bestfitting (#4)
12:00 - 2:00	*Lunch on your own*	

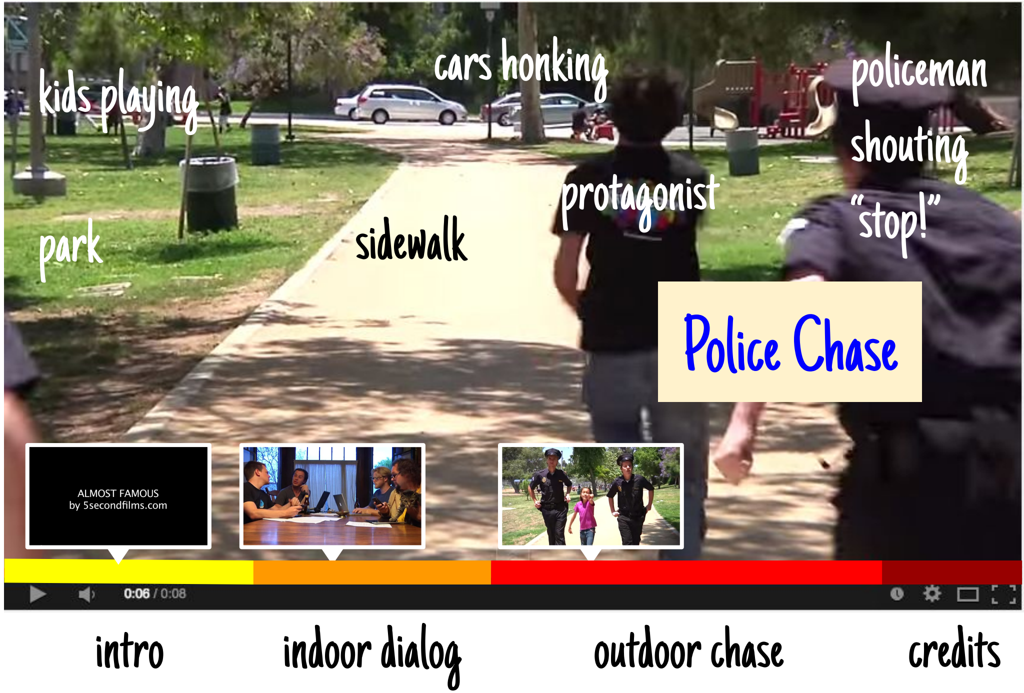
Agenda (Afternoon)

Time	Content	Presenter
Session 3		
2:00 - 2:30	**Invited Talk 3**: Detecting Activities with Less	Cees Snoek
2:30 - 3:00	**Invited Talk 4**: From video-level to fine-grained recognition and retrieval of interactions	Dima Damen
3:00 - 4:00	**Oral Session 2** ● MOD: A Deep Mixture Model with Online Knowledge Distillation for Large Scale Video Temporal Concept Localization ● Cross-Class Relevance Learning for Information Fusion in Temporal Concept Localization ● Noise Learning for Weakly Supervised Segment Classification in Video	● RLin (#3) ● Layer6 AI (#1) ● zhangzhaoyu (#8)
4:00 - 4:30	*Coffee Break*	
Session 4		
4:30 - 6:00	Poster Session	All Accepted Posters

Overview of 2019 YouTube-8M Dataset & Challenge

Joonseok Lee (joonseok@google)
On behalf of the YouTube-8M team

The Multiple Aspects of Video Understanding



Describing the **content:**
what is visible/audible?

Inferring the **central topics:**
what is the story about?

Describing the **structure & style:**
how is the story told?

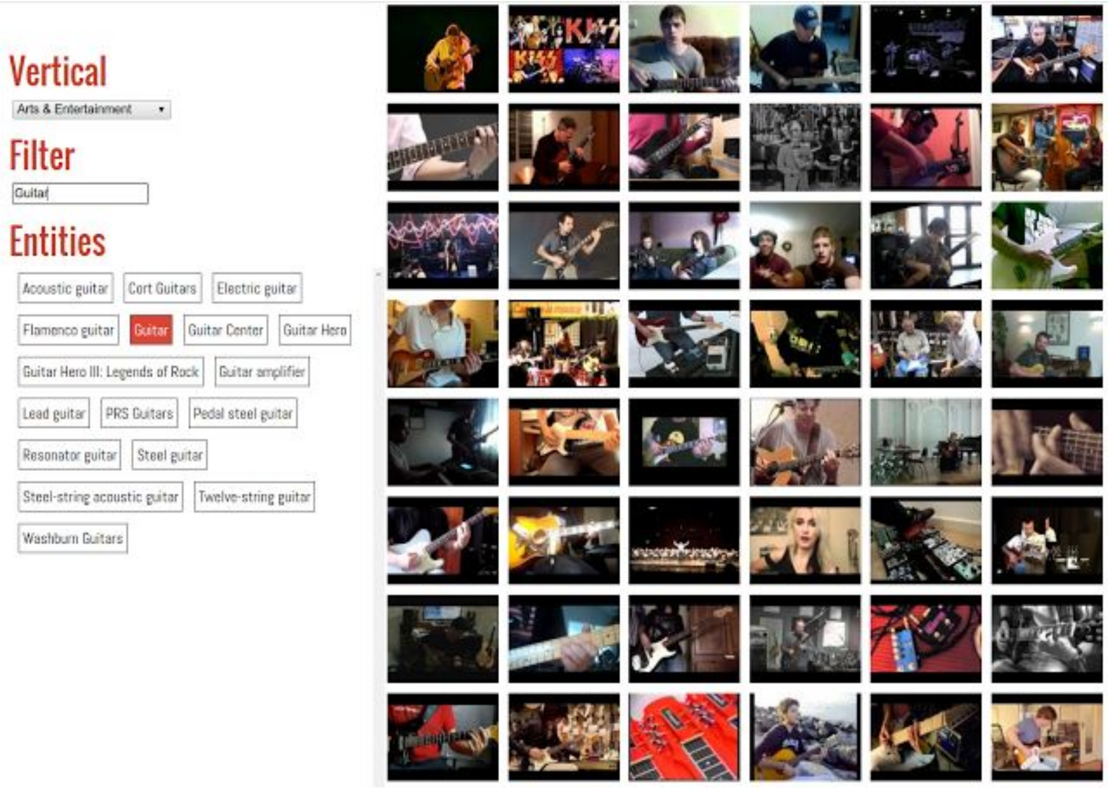
Inferring **creator / viewer intent:**
- **why capture** this video?
- **why watch** this video?

YouTube-8M: Primary Objectives

- Advance the state-of-the-art in Video Understanding
  - By providing a large, free, realistic, labeled video dataset
  - By democratizing research on large-scale video understanding
- Create a representative video annotation benchmark
  - Balancing dataset size and class diversity with training time
  - Key design principles:
    - Preserve the organic distribution as much as possible
    - Make sure all data can fit on a commodity hard disk
    - Make sure a good model can be trained on 1 GPU in < 1 day

The Dataset: YouTube-8M (2018 edition)

- 6.1M videos
- 350,000 hours
- 2.6B audio-visual features
- 3,862 classes
- 3.0 labels/video



Vertical
Arts & Entertainment

Filter
Guitar

Entities

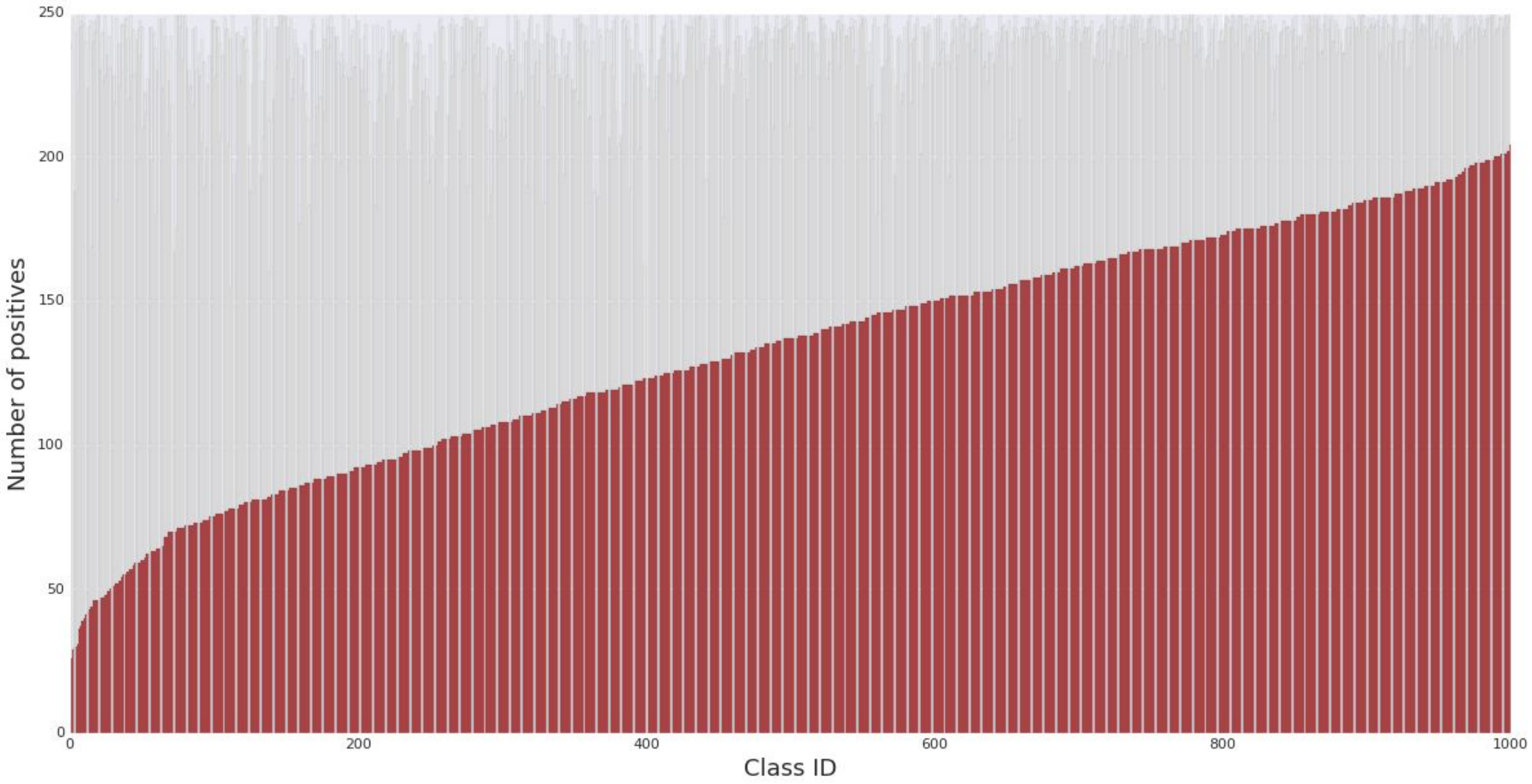
2019 YouTube-8M Challenge: What’s New?

Google AI

YouTube-8M Segments (NEW for 2019)

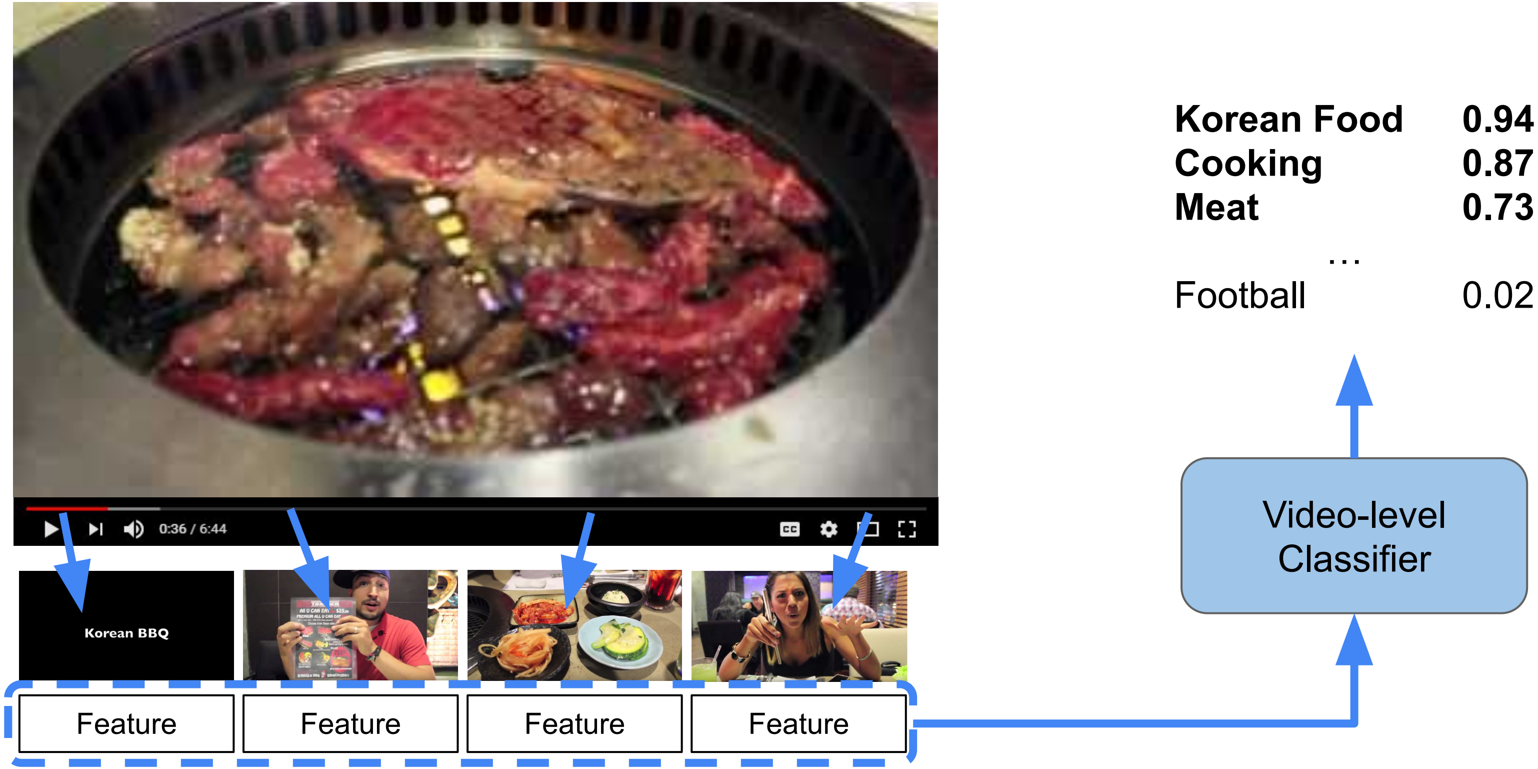
- 1,000 classes (out of 3,862 YT8M vocab) selected based on **temporal-localizability**.
  - E.g., *Typing*, *Squirrel*, *Sunset*, …
    as opposed to *PC Game*, *Concert*, *Football*, ...
- 5 segments/video sampled to label
  - Tried to have **at least one positive** and **one negative** segment.
  - ~80% videos have both positive and negative segments.
- 230K **human-verified** segment labels collected.

YouTube-8M Segments (NEW): Label Distribution



Google AI

Previous Years: Video-level Classification



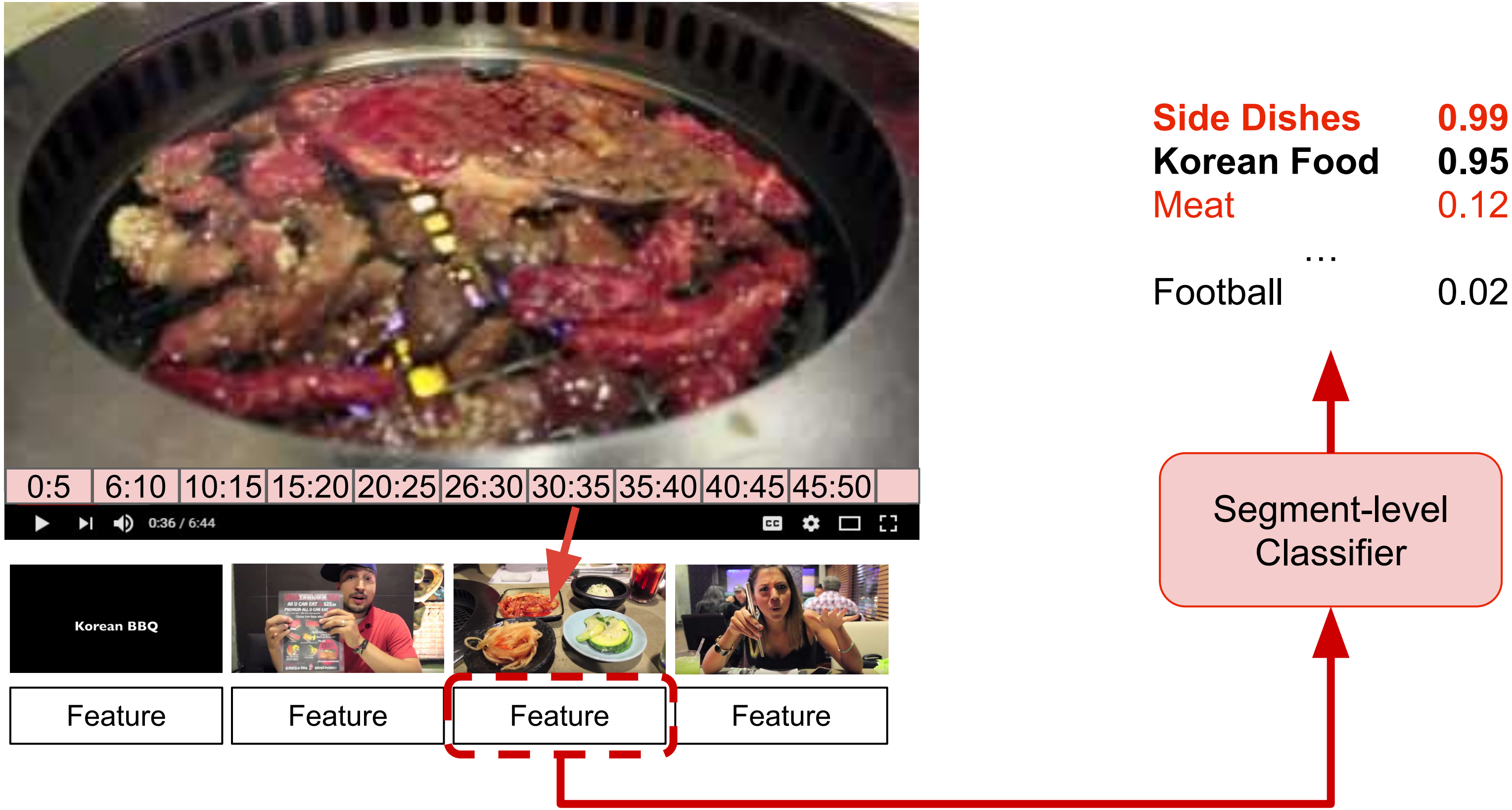
Feature | Feature | Feature | Feature

Video-level Classifier

Korean Food	**0.94**
Cooking	**0.87**
Meat	**0.73**
…	
Football	0.02

Temporal Localization Task

This Year: Segment-level Classification



Segment-level Classifier

Side Dishes	**0.99**
Korean Food	**0.95**
Meat	0.12
…	
Football	0.02

2019 YouTube-8M Challenge Task

- Training data:
  - Frame-level features
    - Visual Inception-V3 bottleneck features extracted from pixels (**1024D**)
    - Audio Resnet-ish bottleneck features extracted from spectrograms (**128D**)
  - Video-level **noisy** labels for 6M+ videos (cover the **main themes** in the video)
  - New in 2019: 5s-long Segment-level **human-verified** labels for 230k+ segments
- Goal:
  - New in 2019: **Predict target segment topics** from the sequence of frame-level features and noisy video-level labels (+some segment-level validation set)
  - Segment topics are from 1,000 entities (subset of 3,862 YT8M vocab)
- New in 2019: Removed model size restriction

Evaluation Metrics

- Mean Average Precision (MAP): Mean per-class AUC of P-R curves

$$mAP = \frac{1}{|E|}\sum_{e} AP(e) = \frac{1}{|E|}\sum_{e=1}^{|E|}\sum_{i=1}^{N} P_{e}(i)\,\Delta R_{e}(i)$$

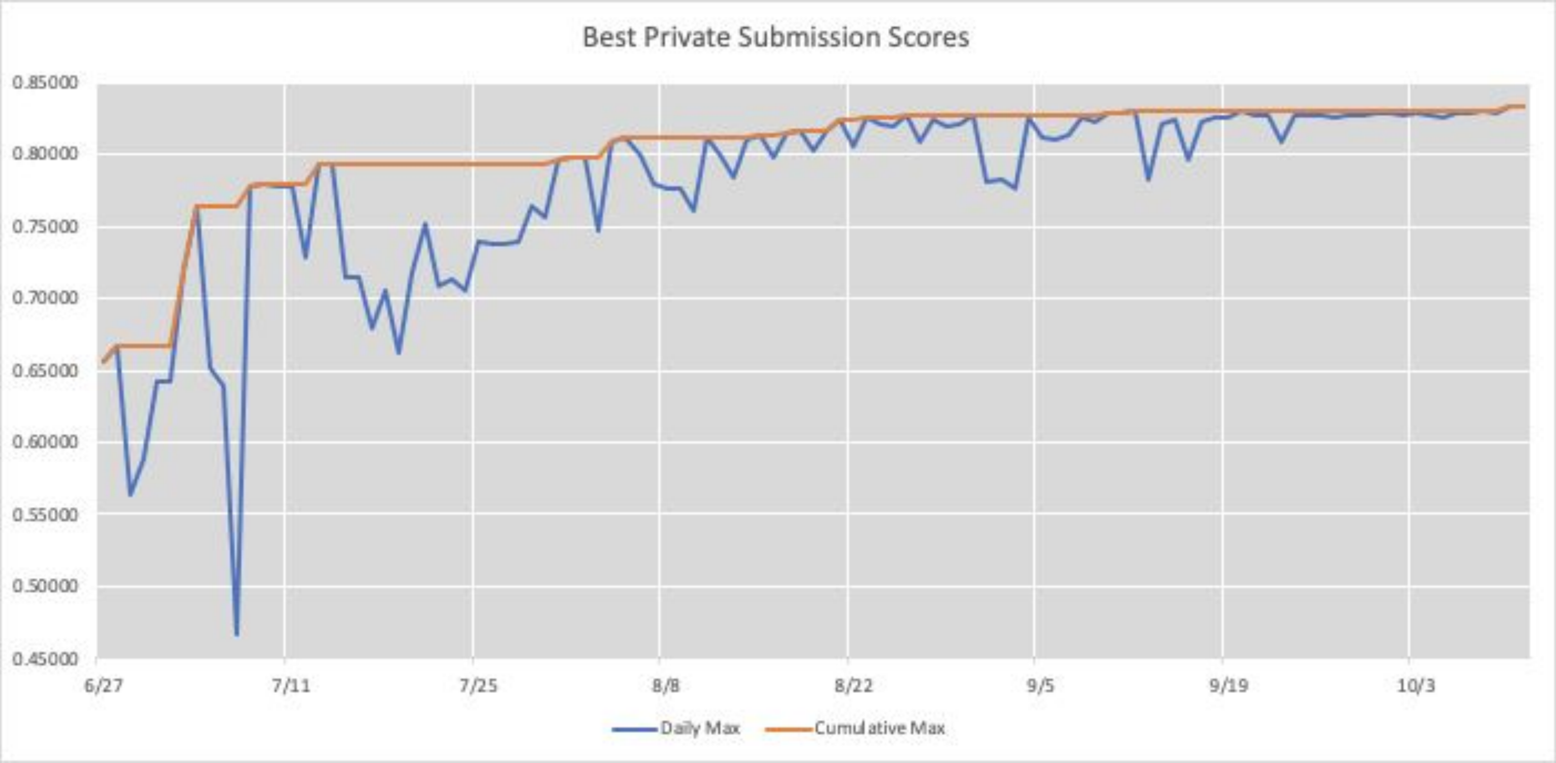
- With this change from global Average Precision (gAP), it is more important to precisely predict **rare classes**. (Each class is equally important regardless of available samples.)

Where were the participants from?

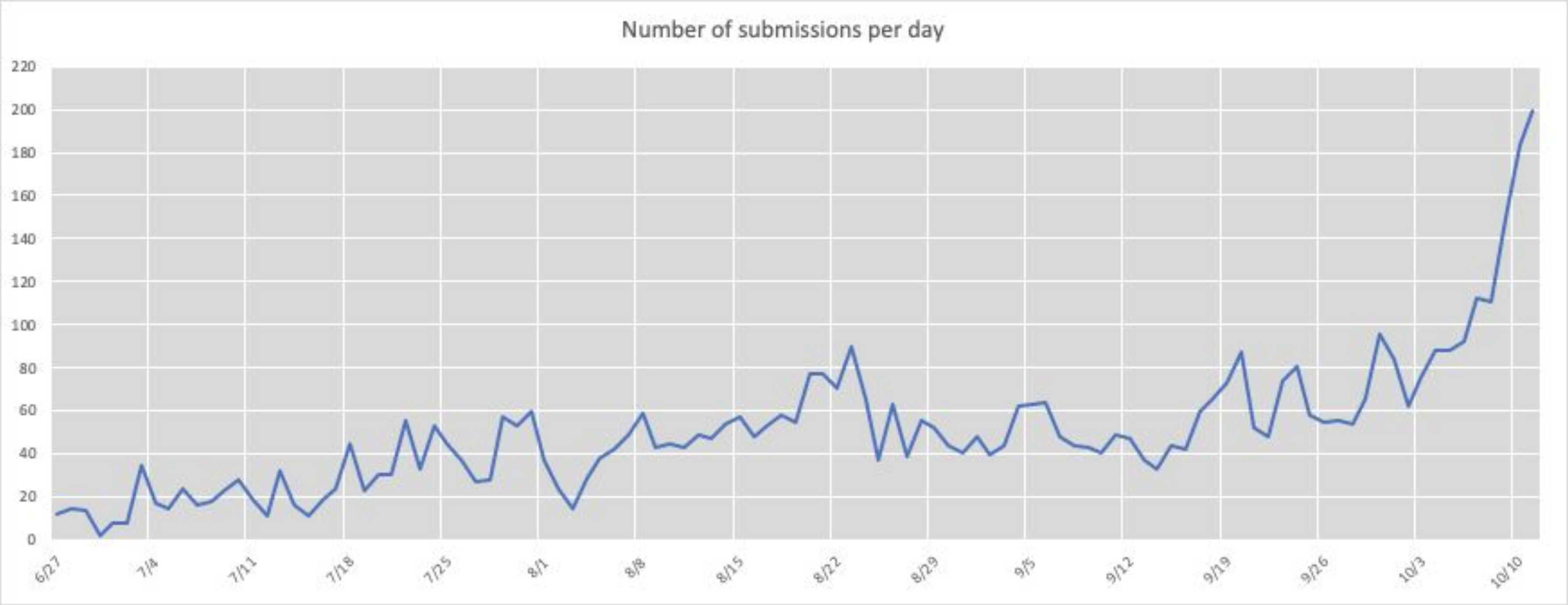
- 283 teams
- 341 competitors
- Participants from 40+ countries
- Total of 3,753 submissions

Country	#Competitors
USA	108
India	46
China PRC	32
Russia	19
Japan	17
Hong Kong	15
France	14
Korea	12
Canada	12
Taiwan	9
UK	7
Ukraine	6
Pakistan	6
Germany	6
Sweden	4
Saudi Arabia	4
Turkey	3
Thailand	3

Competition Progression

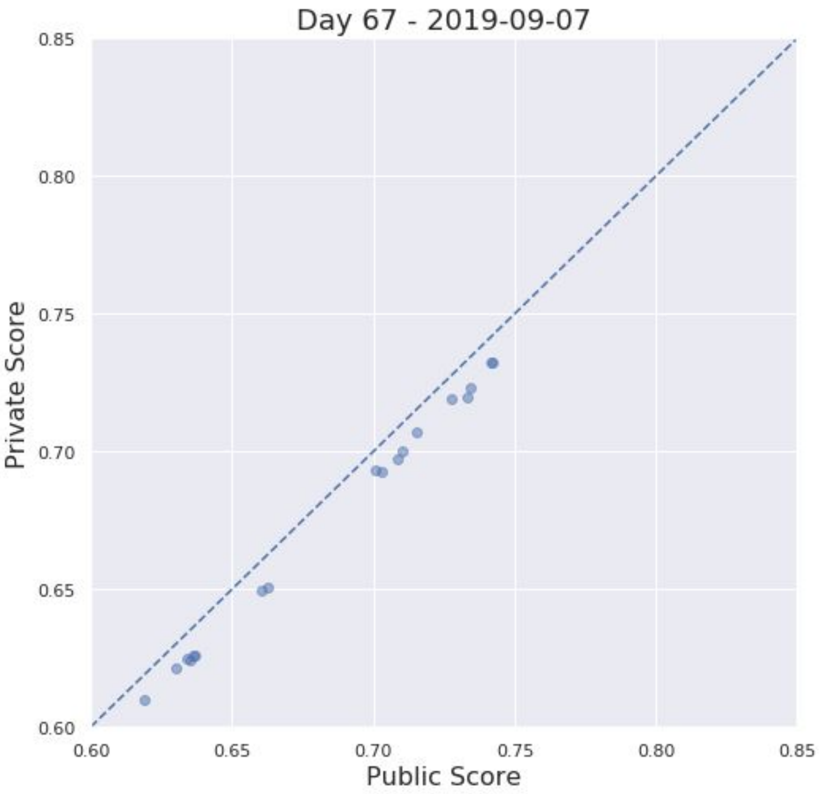


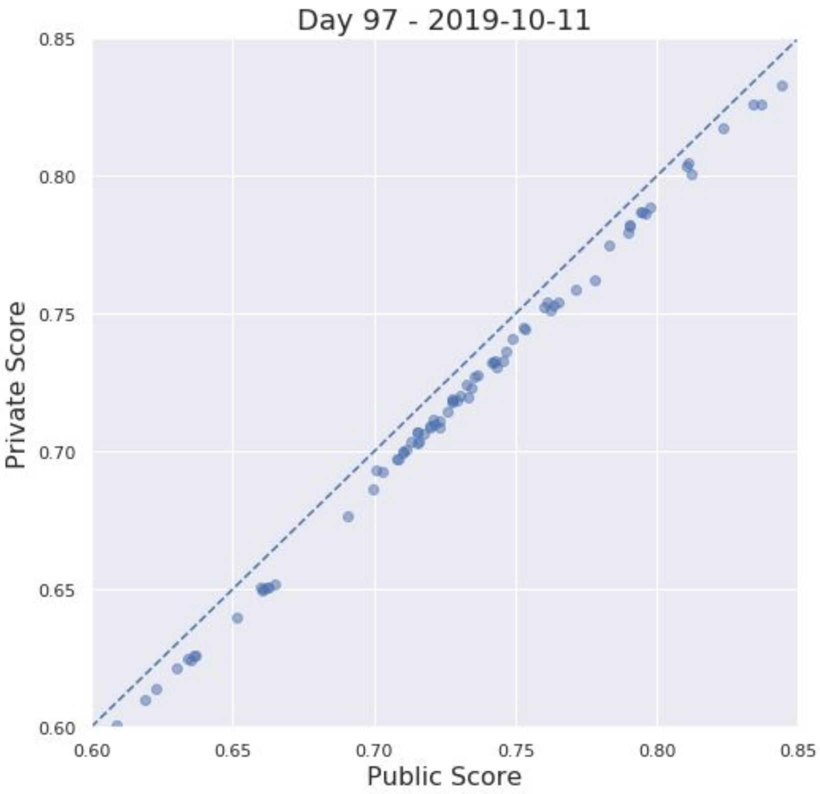
Competition Progression



Google AI

Did the models overfit the Public Test data?





Google AI

Logistics

Agenda (Morning)

Time	Content	Presenter
9:00 - 9:05	Opening Remarks	Paul Natsev
9:05 - 9:20	**Overview of 2019 YouTube-8M Dataset & Challenge**	Joonseok Lee
Session 1		
9:20 - 9:50	**Invited Talk 1**: Human action recognition and the Kinetics dataset	Jitendra Malik
9:50 - 10:20	**Invited Talk 2**: Learning from Narrated Videos	Jean-Baptiste Alayrac
10:20 - 10:40	*Coffee Break*	
Session 2		
10:40 - 11:00	MediaPipe: A framework for building perception pipelines	Chris McClanahan
11:00 - 12:00	**Oral Session 1** ● Logistic Regression is Still Alive and Effective:The 3rd YouTube 8M challenge solution of the IVUL-KAUST team ● Multi-attention Networks for Temporal Localization of Video-level Labels ● A segment-level classification solution to the 3rd YouTube-8M Video Understanding Challenge	● IVUL-KAUST (#11) ● Locust (#13) ● bestfitting (#4)
12:00 - 2:00	*Lunch on your own*	

Agenda (Afternoon)

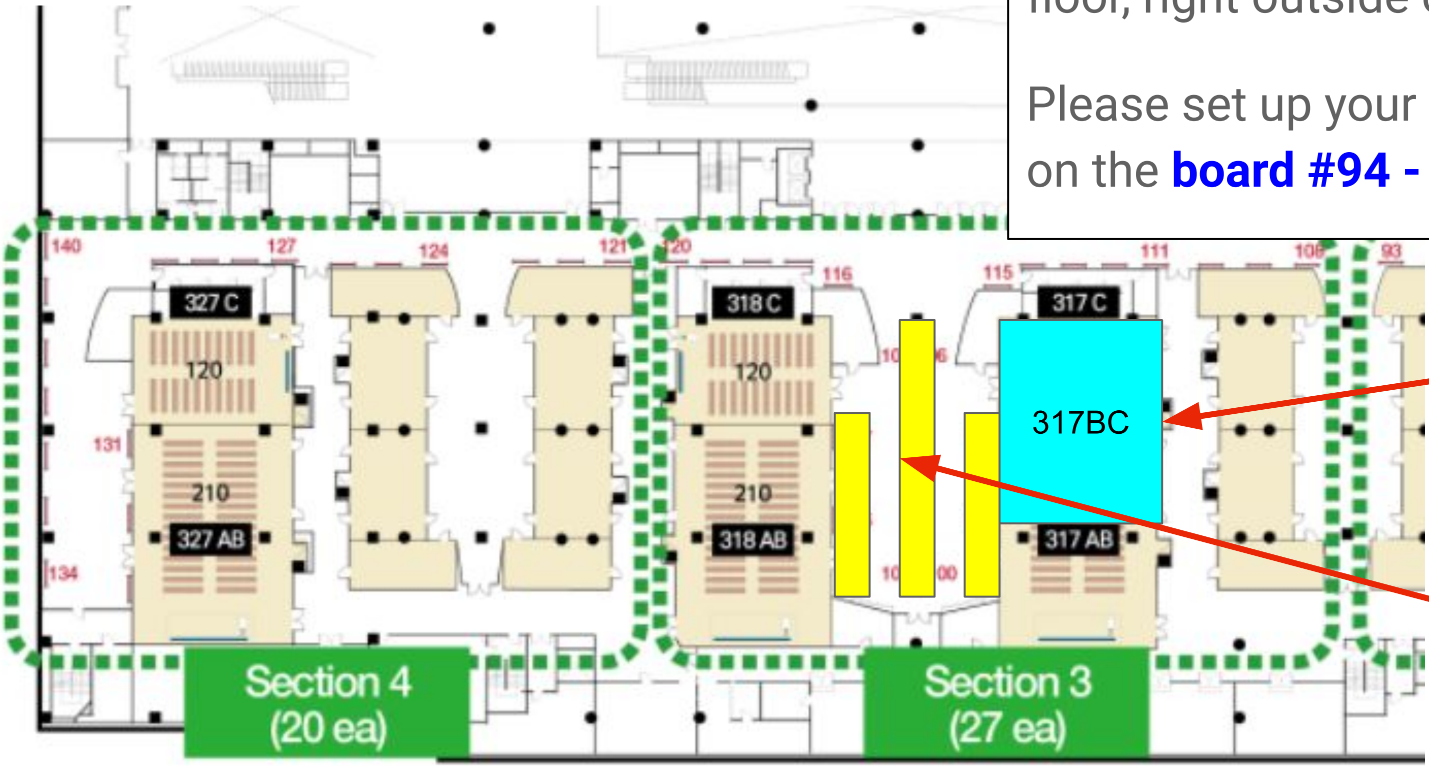
Time	Content	Presenter
Session 3		
2:00 - 2:30	**Invited Talk 3**: Detecting Activities with Less	Cees Snoek
2:30 - 3:00	**Invited Talk 4**: From video-level to fine-grained recognition and retrieval of interactions	Dima Damen
3:00 - 4:00	**Oral Session 2** ● MOD: A Deep Mixture Model with Online Knowledge Distillation for Large Scale Video Temporal Concept Localization ● Cross-Class Relevance Learning for Information Fusion in Temporal Concept Localization ● Noise Learning for Weakly Supervised Segment Classification in Video	● RLin (#3) ● Layer6 AI (#1) ● zhangzhaoyu (#8)
4:00 - 4:30	*Coffee Break*	
Session 4		
4:30 - 6:00	Poster Session	All Accepted Posters

Poster Session Location

We are on **3rd** floor now. (317BC)

Poster session will be on the same floor, right outside of our room.

Please set up your poster on the **board #94 - 113 after 1pm**.



We are here now.

Poster session here
(4:30 - 6:00)
Board 94 - 113

MediaPipe

A framework for building perception pipelines

Chris McClanahan, Google Research

What is MediaPipe?

MediaPipe is Google's **cross-platform** framework
for building **perception pipelines**

Widely used at Google in **research & products** to process and analyze video, audio and sensor data

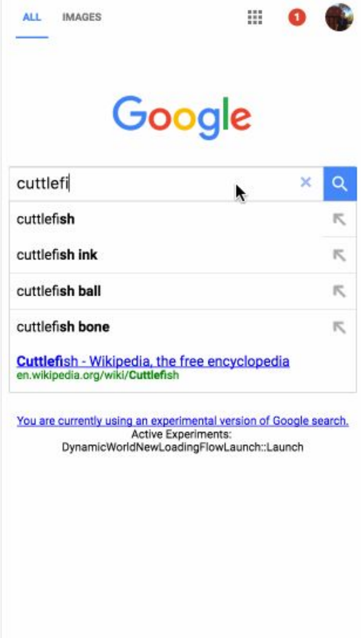
- Dataset preparation pipelines for ML training
- ML inference pipelines
- Media processing pipelines

mediapipe.dev

MediaPipe in Production

Mobile:
Visual Search, Lens

Server:
Video Previews



ALL IMAGES

Google

cuttlefi

cuttlefish

cuttlefish ink

cuttlefish ball

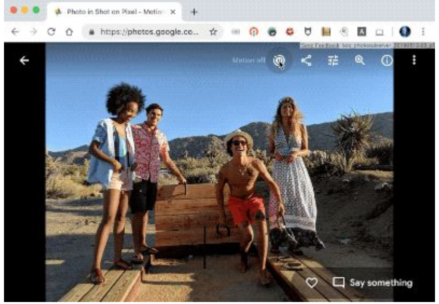
cuttlefish bone

Cuttlefish - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Cuttlefish

You are currently using an experimental version of Google search.
Active Experiments:
DynamicWorldNewLoadingFlowLaunch::Launch

Cross-platform:
Motion Photos

Android



Server / Browser

AR:
YouTube, ARCore, Duo

YT8M Feature Extraction & Model Inference with MediaPipe

New Tools for YT8M

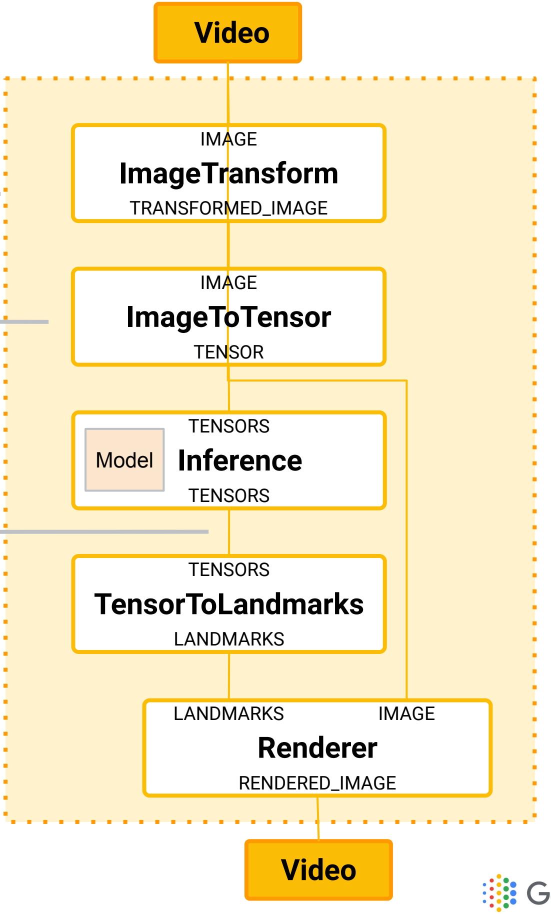
- Feature extraction / Dataset preparation pipeline:
  - Local video path in TFRecord in -> features in TFRecord out.
- Model inference pipelines:
  - Local video path in TFRecord + features in TFRecord in -> annotated video out
  - YT8M features in TFRecord in -> labels out
  - Web Interface

MediaPipe Concepts

A MediaPipe **Graph** represents a **perception pipeline**

Each node in the pipeline is a MediaPipe **Calculator**

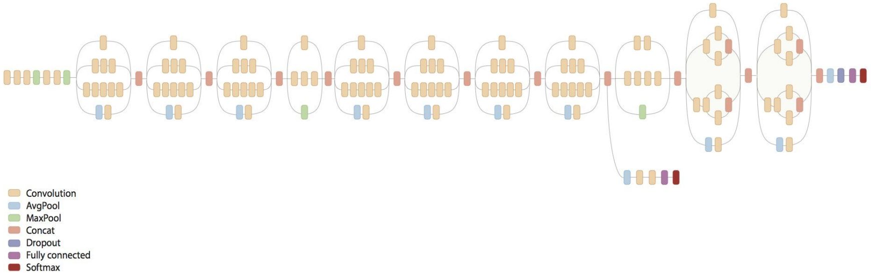
A pair of nodes are connected by a **Stream**, which carries a sequence of **Packets** with ascending timestamps

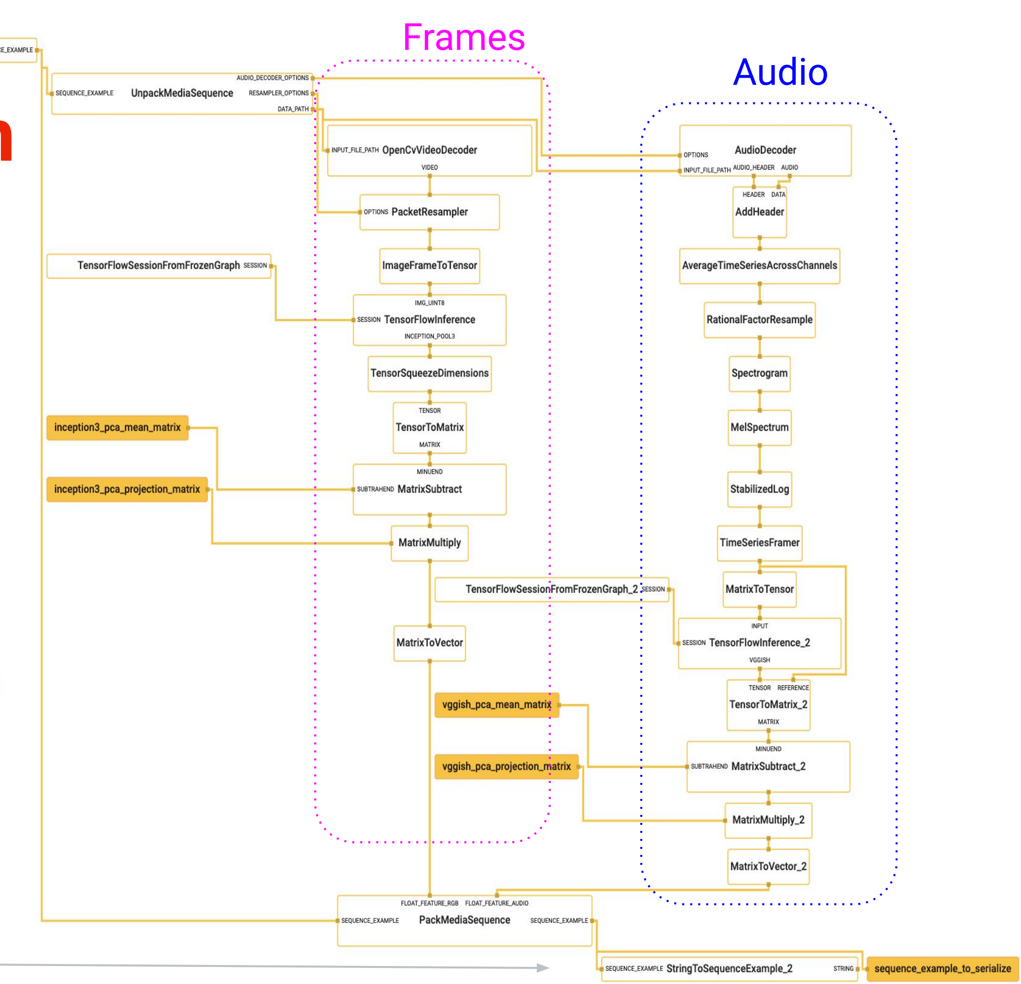


YT8M Feature Extraction

Input: video path in TFRecord

```
Example command:
$ ./extract_yt8m_features \
    --calculator_graph=feature_extraction.pbtxt \
    --input_side_packets=/tmp/input.tfrecord \
    --output_side_packets=/tmp/output.tfrecord
```





Output: features in TFRecord

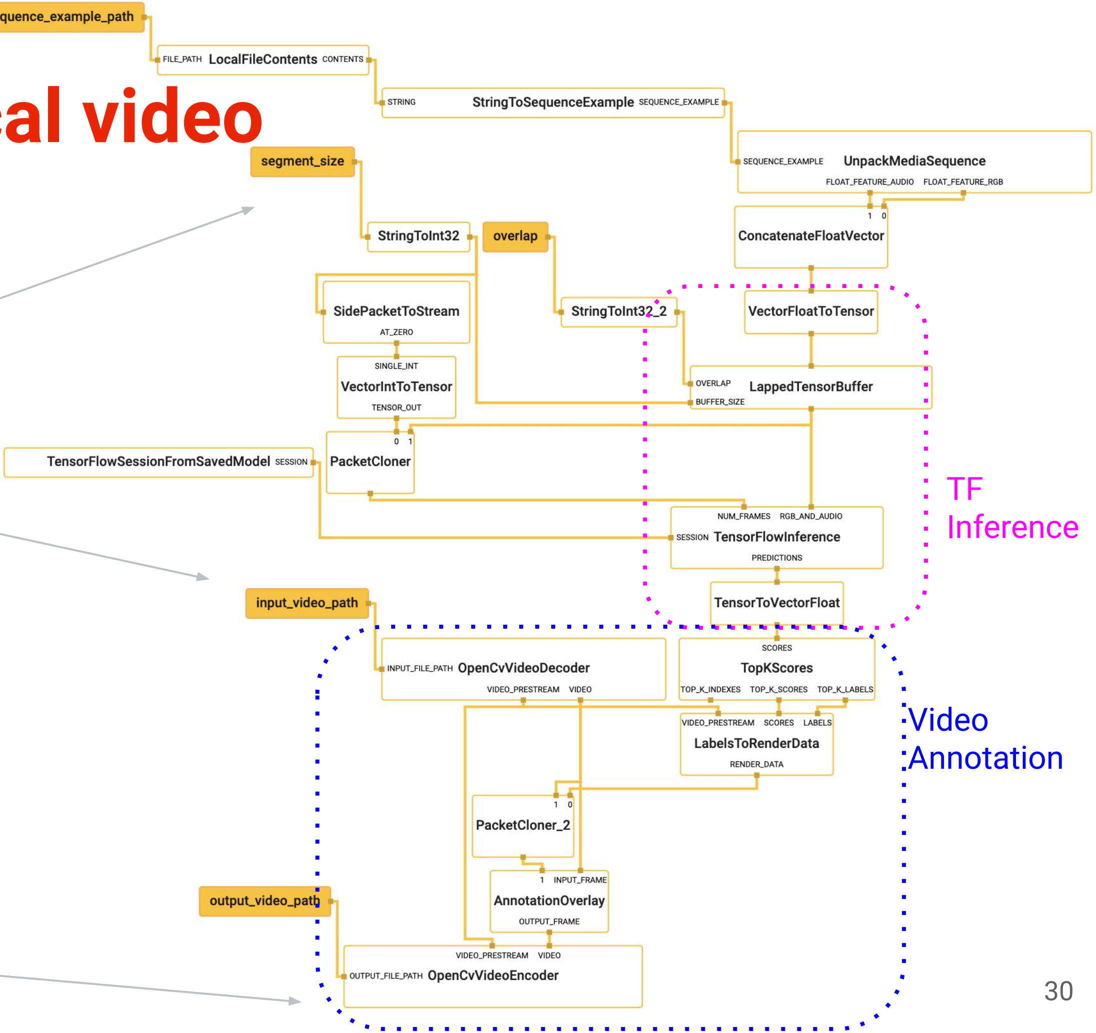
Model Inference - Local video

Input: features in TFRecord
(generated by feature extraction)

Input: segment size & overlap

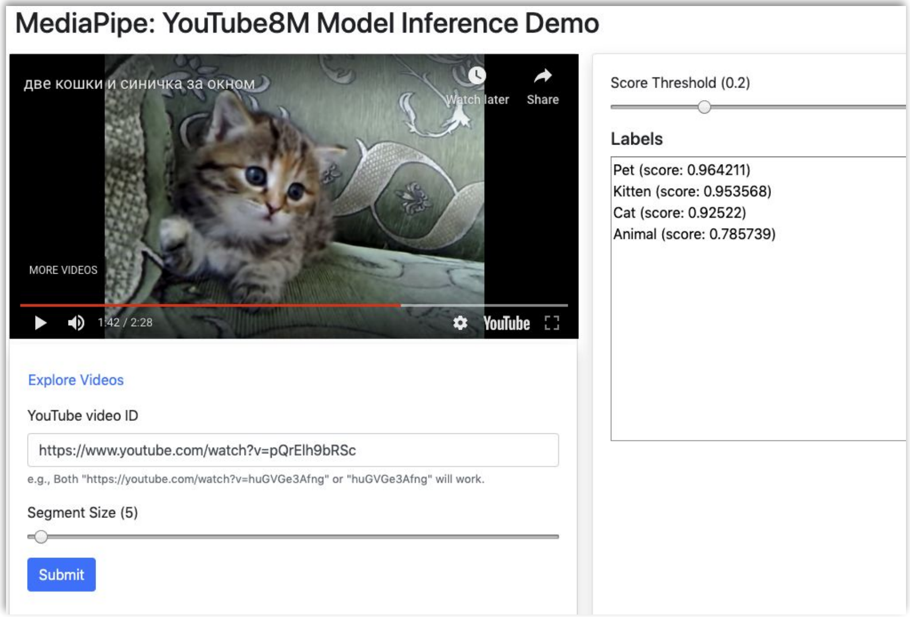
Input: path to video file
(for rendering video)

Output: annotated video



Model Inference - Web Interface

- Easy way to test your models on the dataset
- Automatically looks up YT8M id & TfRecord
- See segment labels synced with video
- **Live demo** using baseline model…
  - Deep Bag-of-Frame (DBoF)



MediaPipe: YouTube8M Model Inference Demo

Explore Videos

YouTube video ID

https://www.youtube.com/watch?v=pQrElh9bRSc

e.g., Both "https://youtube.com/watch?v=huGVGe3Afng" or "huGVGe3Afng" will work.

Segment Size (5)

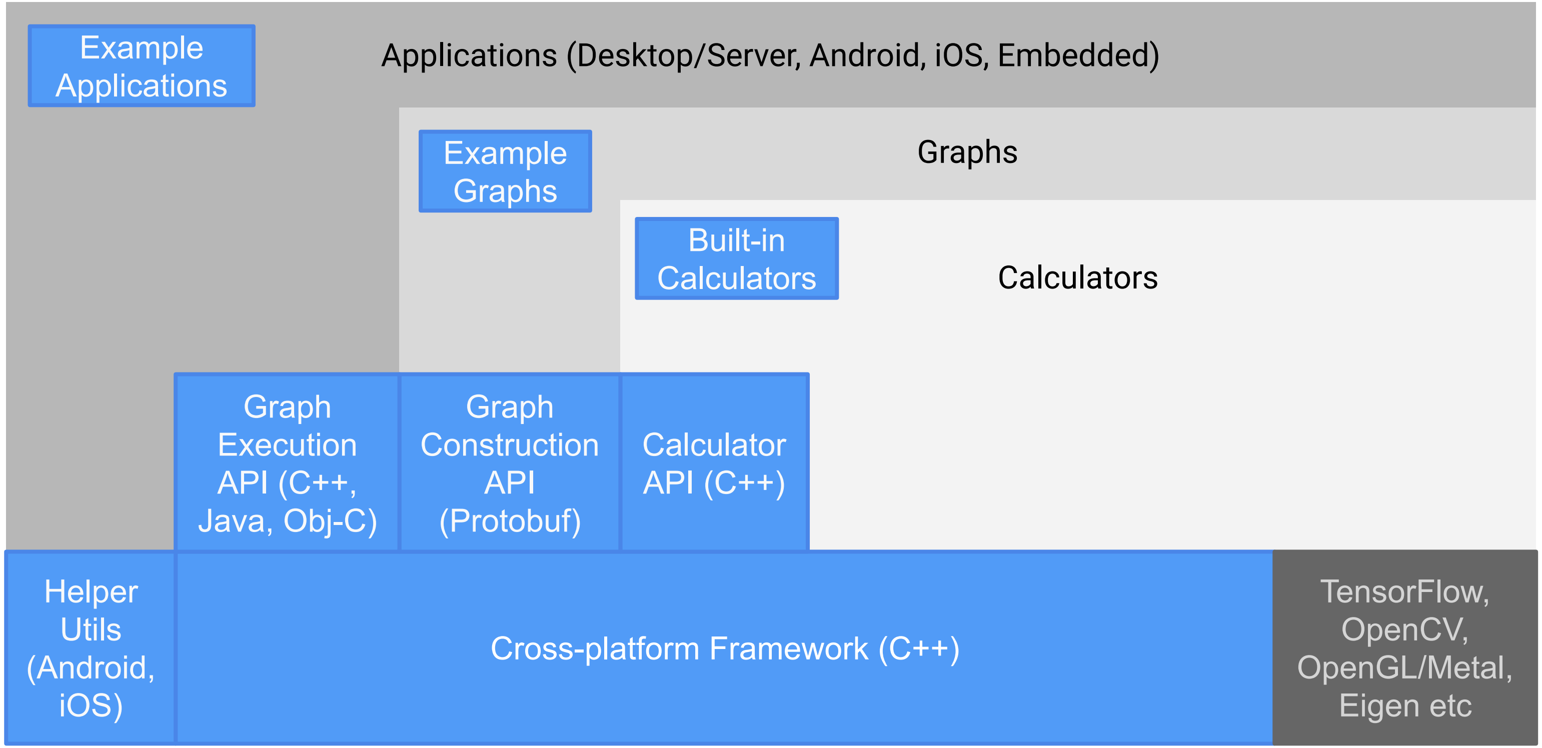
Submit

Score Threshold (0.2)

Labels

Pet (score: 0.964211)
Kitten (score: 0.953568)
Cat (score: 0.92522)
Animal (score: 0.785739)

MediaPipe Tech Stack



Example Applications

Applications (Desktop/Server, Android, iOS, Embedded)

Example Graphs

Graphs

Built-in Calculators

Calculators

Graph Execution API (C++, Java, Obj-C)

Graph Construction API (Protobuf)

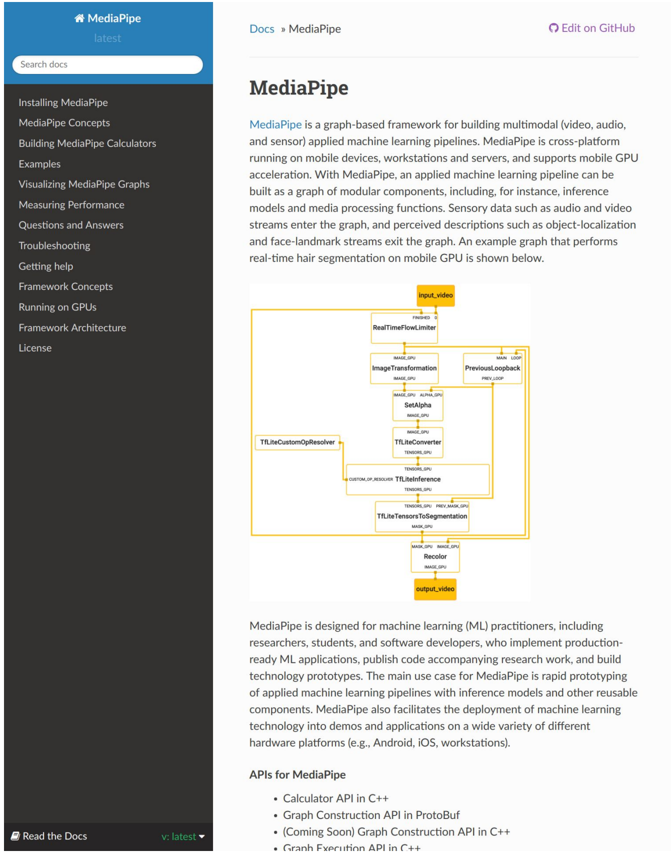
Calculator API (C++)

Helper Utils (Android, iOS)

Cross-platform Framework (C++)

TensorFlow, OpenCV, OpenGL/Metal, Eigen etc

docs.mediapipe.dev



MediaPipe

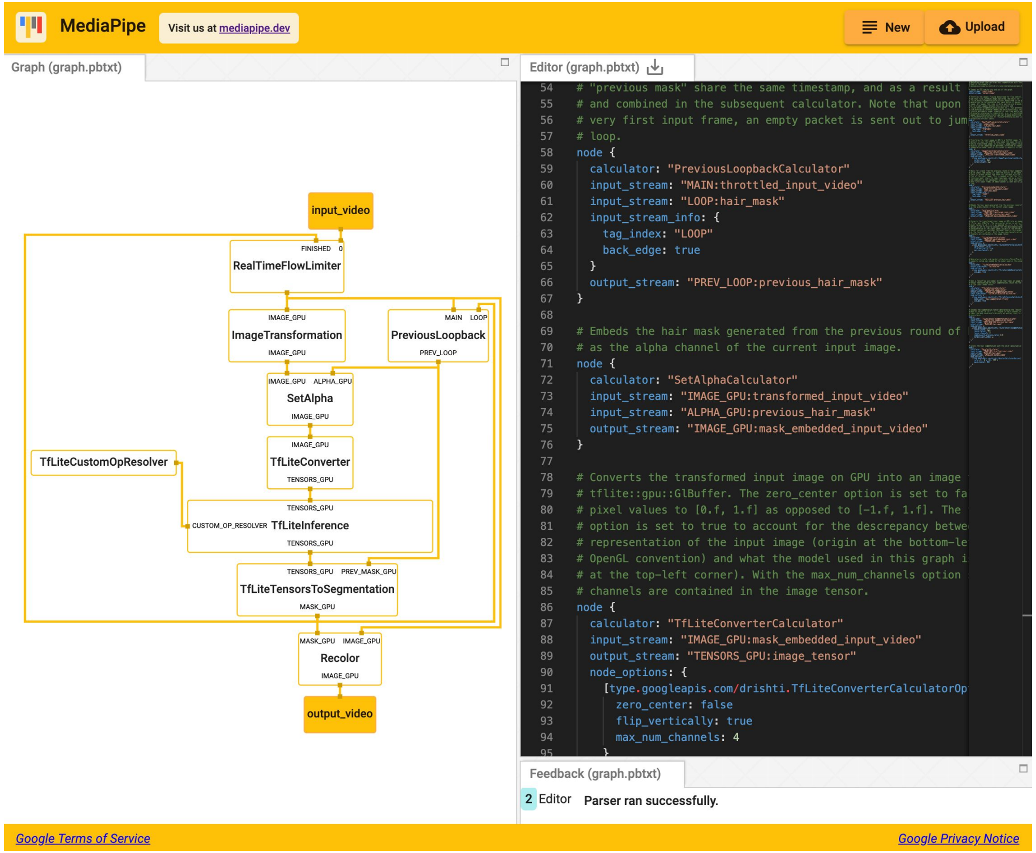
MediaPipe is a graph-based framework for building multimodal (video, audio, and sensor) applied machine learning pipelines. MediaPipe is cross-platform running on mobile devices, workstations and servers, and supports mobile GPU acceleration. With MediaPipe, an applied machine learning pipeline can be built as a graph of modular components, including, for instance, inference models and media processing functions. Sensory data such as audio and video streams enter the graph, and perceived descriptions such as object-localization and face-landmark streams exit the graph. An example graph that performs real-time hair segmentation on mobile GPU is shown below.

MediaPipe is designed for machine learning (ML) practitioners, including researchers, students, and software developers, who implement production-ready ML applications, publish code accompanying research work, and build technology prototypes. The main use case for MediaPipe is rapid prototyping of applied machine learning pipelines with inference models and other reusable components. MediaPipe also facilitates the deployment of machine learning technology into demos and applications on a wide variety of different hardware platforms (e.g., Android, iOS, workstations).

APIs for MediaPipe

- Calculator API in C++
- Graph Construction API in ProtoBuf
- (Coming Soon) Graph Construction API in C++
- Graph Execution API in C++

viz.mediapipe.dev



```
# "previous mask" share the same timestamp, and as a result
# and combined in the subsequent calculator. Note that upon
# very first input frame, an empty packet is sent out to jum
# loop.
node {
  calculator: "PreviousLoopbackCalculator"
  input_stream: "MAIN:throttled_input_video"
  input_stream: "LOOP:hair_mask"
  input_stream_info: {
    tag_index: "LOOP"
    back_edge: true
  }
  output_stream: "PREV_LOOP:previous_hair_mask"
}

# Embeds the hair mask generated from the previous round of
# as the alpha channel of the current input image.
node {
  calculator: "SetAlphaCalculator"
  input_stream: "IMAGE_GPU:transformed_input_video"
  input_stream: "ALPHA_GPU:previous_hair_mask"
  output_stream: "IMAGE_GPU:mask_embedded_input_video"
}

# Converts the transformed input image on GPU into an image
# tflite::gpu::GlBuffer. The zero_center option is set to fa
# pixel values to [0.f, 1.f] as opposed to [-1.f, 1.f]. The
# option is set to true to account for the descrepancy betwe
# representation of the input image (origin at the bottom-le
# OpenGL convention) and what the model used in this graph i
# at the top-left corner). With the max_num_channels option
# channels are contained in the image tensor.
node {
  calculator: "TfLiteConverterCalculator"
  input_stream: "IMAGE_GPU:mask_embedded_input_video"
  output_stream: "TENSORS_GPU:image_tensor"
  node_options: {
    [type.googleapis.com/drishti.TfLiteConverterCalculatorOp
      zero_center: false
      flip_vertically: true
      max_num_channels: 4
    }
```

2 Editor Parser ran successfully.

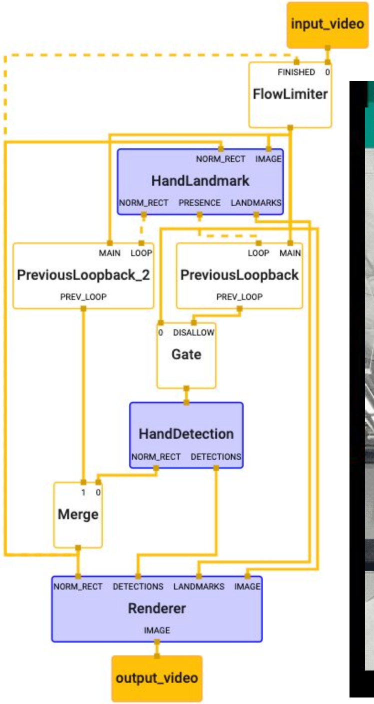
Desktop/Server Examples

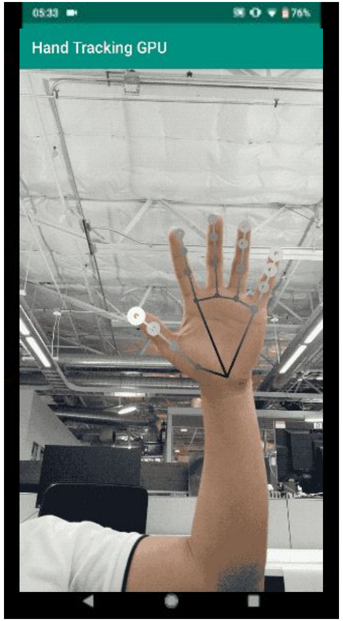
- YouTube 8M feature extraction and model inference
- Data preparation using MediaSequence (e.g., for DeepMind Kinetics i3d)
- Face detection
- Object detection
- Hand tracking

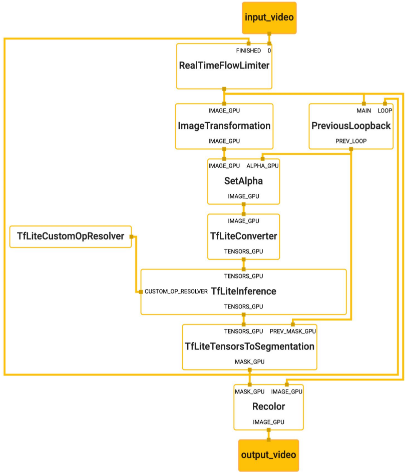
mediapipe.dev

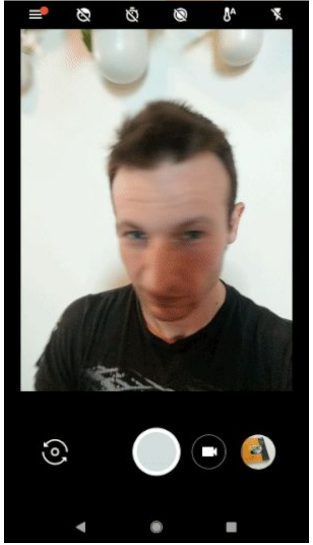
Mobile Examples

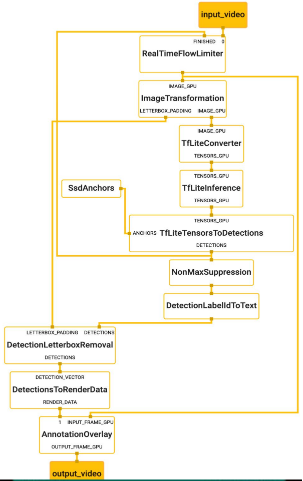
mediapipe.dev

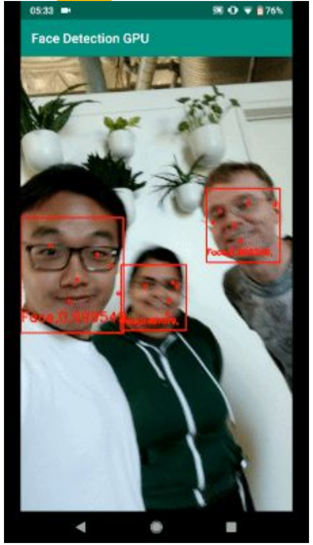


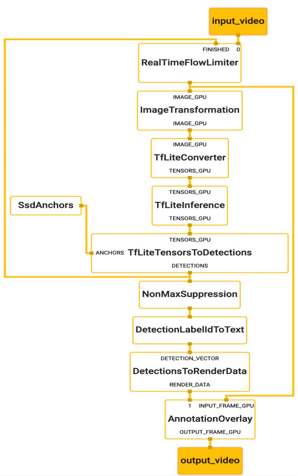


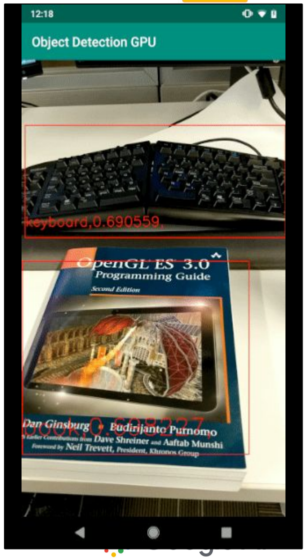








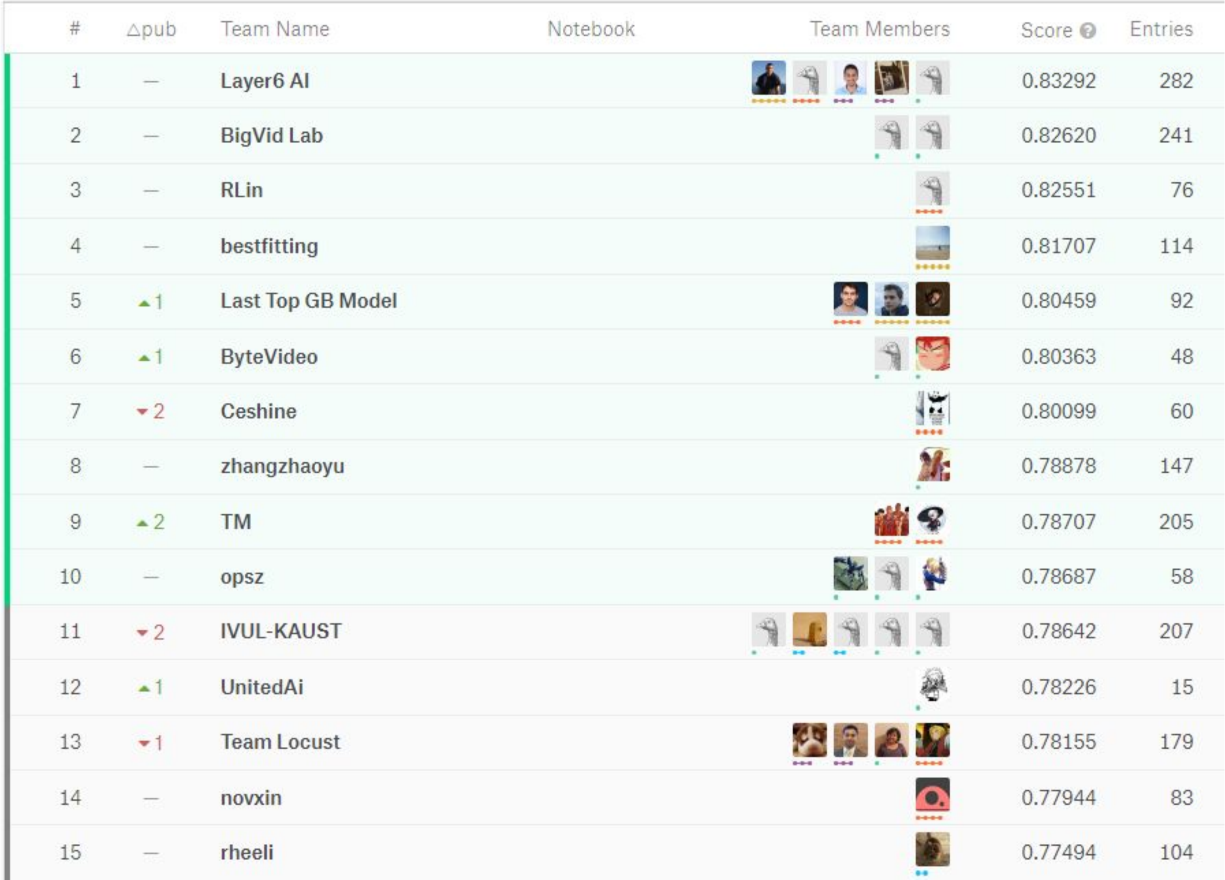



12:18
Object Detection GPU
keyboard,0.690559
OpenGL ES 3.0
Programming Guide
Second Edition
Dan Ginsburg Budirijanto Purnomo
Dave Shreiner and Aaftab Munshi
Foreword by Neil Trevett, President, Khronos Group

Thank you for your attention.

Closing Remarks

Final Private Leaderboard



#	△pub	Team Name	Notebook	Team Members	Score	Entries
1	—	Layer6 AI			0.83292	282
2	—	BigVid Lab			0.82620	241
3	—	RLin			0.82551	76
4	—	bestfitting			0.81707	114
5	▲ 1	Last Top GB Model			0.80459	92
6	▲ 1	ByteVideo			0.80363	48
7	▼ 2	Ceshine			0.80099	60
8	—	zhangzhaoyu			0.78878	147
9	▲ 2	TM			0.78707	205
10	—	opsz			0.78687	58
11	▼ 2	IVUL-KAUST			0.78642	207
12	▲ 1	UnitedAi			0.78226	15
13	▼ 1	Team Locust			0.78155	179
14	—	novxin			0.77944	83
15	—	rheeli			0.77494	104

The Winner: Layer6 AI

- Members
  - Junwei Ma (Layer6 AI)
  - Satya Krishna Gorti (Layer6 AI)
  - Maksims Volkovs (Layer6 AI)
  - Ilya Stanevich (Layer6 AI)
  - Guangwei Yu (Layer6 AI)

- Score on public evalset: 0.84429 (#1)
  Score on private evalset: 0.83292 (#1)

Special Thanks to

- #3: Team **RLin** (#3 in 2018)
  - Rongcheng Lin (University of North Carolina at Charlotte)
  - Jing Xiao (University of North Carolina at Charlotte)
  - Jianping Fan (University of North Carolina at Charlotte)

- #5: Team **Last Top GB Model** (#1 in 2018, #5 in 2017)
  - Miha Skalic (University Pompeu Fabra)
  - Mikel Bober-Irizar (Royal Grammar School Guildford)
  - David Austin (Intel)

Acknowledgments

General Chairs

Paul Natsev

Rahul Sukthankar

Cordelia Schmid

Program Chairs

Joonseok Lee

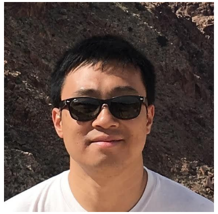
George Toderici

kaggle

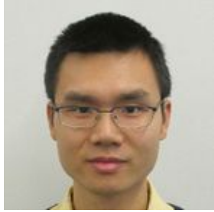
Julia Elliott

Walter Reade

Challenge Organizers


Ke Chen

Nisarg Kothari

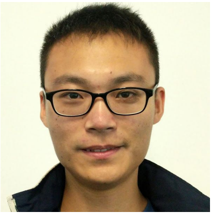

Hanhan Li

Joe Ng

Sobhan Naderi Parizi

David Ross

Javier Snaider


Zheng Xu

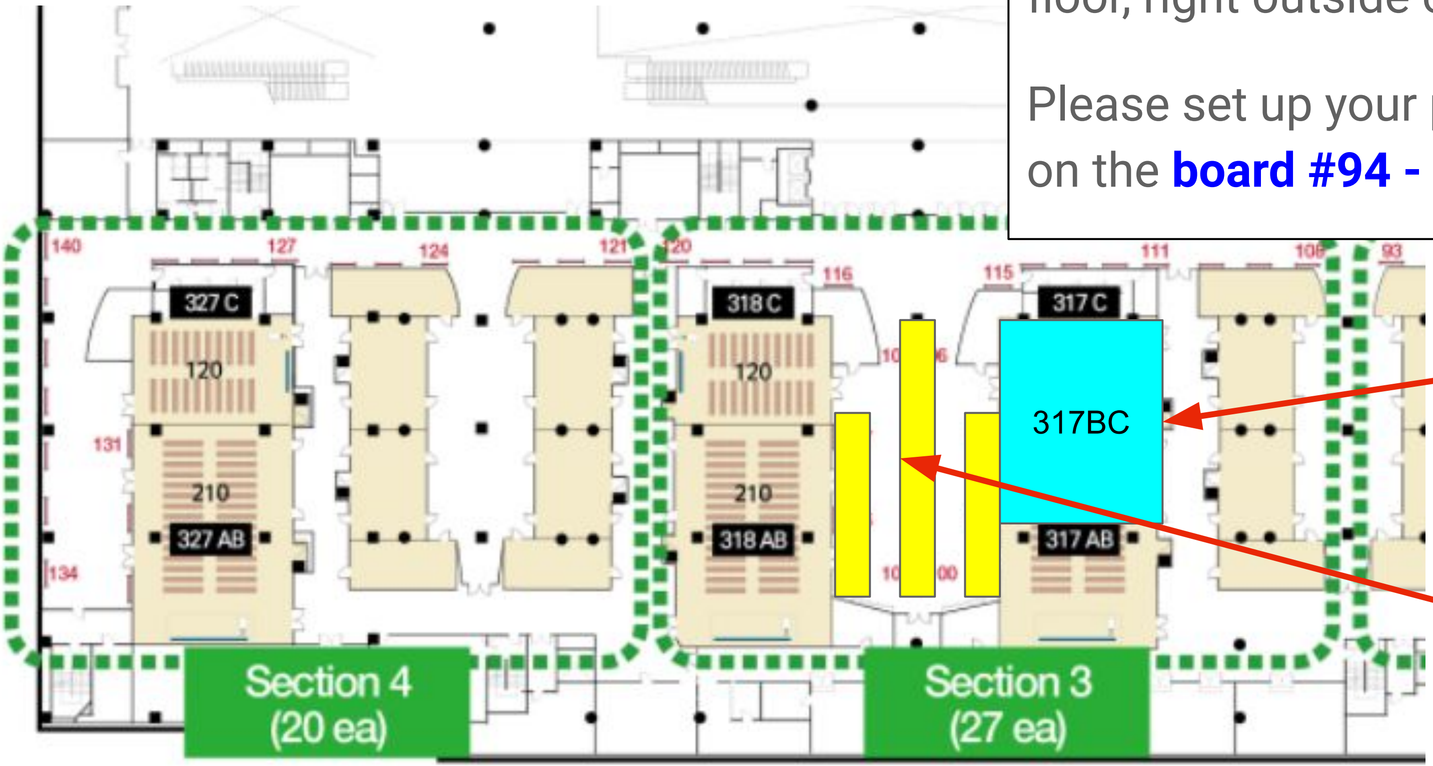
+ ***Big thanks to all speakers and participants!!***

Poster Session Location

We are on **3rd** floor now. (317BC)

Poster session will be on the same floor, right outside of our room.

Please set up your poster on the **board #94 - 113 now**.



We are here now.

Poster session here
(4:30 - 6:00)
Board 94 - 113

Thanks again for participation.

Ideas or suggestions are welcome for the competition!